remote sensing

MDPI

Article

Monitoring Anthropogenically Disturbed Parcels with Soil Erosion Dynamics Change Based on an Improved SegFormer

Zhenqiang Li [1], Jialin Li [1,*], Jie Li [2], Zhangxuan Li [1], Kuncheng Jiang [1], Yuyang Ma [1] and Chuli Hu [1]

[1] National Engineering Research Center for Geographic Information System, School of Geography and Information Engineering, Engineering Research Center of Natural Resource Information Management and Digital Twin Engineering Software, Ministry of Education, China University of Geosciences (Wuhan), Wuhan 430074, China; lizhenqiang@cug.edu.cn (Z.L.); lizhangxuan@cug.edu.cn (Z.L.); 1202310742@cug.edu.cn (K.J.); yuyangma@cug.edu.cn (Y.M.); huchl@cug.edu.cn (C.H.)

[2] School of Computer Science, China University of Geosciences, Wuhan 430074, China; lijie_gis@cug.edu.cn

* Correspondence: lijialin@cug.edu.cn

Abstract: Amidst burgeoning socioeconomic development, anthropogenic activities have exacerbated soil erosion. This erosion, characterized by its brief duration, high frequency, and considerable environmental degradation, presents a major challenge to ecological systems. Therefore, it is imperative to regulate and remediate erosion–prone, anthropogenically disturbed parcels, with dynamic change detection (CD) playing a crucial role in enhancing management efficiency. Currently, traditional methods for change detection, such as field surveys and visual interpretation, suffer from time inefficiencies, complexity, and high resource consumption. Meanwhile, despite advancements in remote sensing technology that have improved the temporal and spatial resolution of images, the complexity and heterogeneity of terrestrial cover types continue to limit large–scale dynamic monitoring of anthropogenically disturbed soil erosion parcels (ADPSE) using remote sensing techniques. To address this, we propose a novel ISegFormer model, which integrates the SegFormer network with a pseudo–residual multilayer perceptron (PR–MLP), cross–scale boundary constraint module (CSBC), and multiscale feature fusion module (MSFF). The PR–MLP module improves feature extraction by capturing spatial contextual information, while the CSBC module enhances boundary prediction through high– and low–level semantic guidance. The MSFF module fuses multiscale features with attention mechanisms, boosting segmentation precision for diverse change types. Model performance is evaluated using metrics, such as precision, recall, F1–score, intersection over union (IOU), and mean intersection over union (mIOU). The results demonstrate that our improved model performs exceptionally well in dynamic monitoring tasks for ADPSE. Compared to five other models, our model achieved an mIOU of 72.34% and a Macro–F1 score of 83.55% across twelve types of ADPSE changes, surpassing the other models by 1.52–2.48% in mIOU and 2.25–3.64% in Macro–F1 score. This work provides a theoretical and methodological foundation for policy–making in soil and water conservation departments.

Keywords: change detection; anthropogenic activities; dynamic monitoring; segformer; remote sensing; deep learning

check for updates

Citation: Li, Z.; Li, J.; Li, J.; Li, Z.; Jiang, K.; Ma, Y.; Hu, C. Monitoring Anthropogenically Disturbed Parcels with Soil Erosion Dynamics Change Based on an Improved SegFormer. *Remote Sens.* **2024**, *16*, 4494. https://doi.org/10.3390/rs16234494

Academic Editor: Dimitrios D. Alexakis

Received: 15 October 2024
Revised: 20 November 2024
Accepted: 26 November 2024
Published: 29 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a key component of global environmental change research, soil erosion is a primary factor contributing to soil quality degradation and land productivity decline. Sediment erosion and nutrient loss in soils cause severe damage to the ecological environment of soil and water [1]. Natural and anthropogenic factors can lead to soil erosion, with human activities often being one of the primary causes of severe soil degradation [2–4]. Natural factors contributing to soil erosion include weather conditions, topographical features, land use patterns, and vegetation cover, while anthropogenic factors encompass engineering

projects, agricultural and forestry development, and road construction. China is one of the countries most severely affected by soil erosion. According to the 2020 national dynamic monitoring data on soil and water loss, the area of soil erosion in China has reached 2,692,700 km^2, making soil erosion a pressing environmental issue that urgently needs to be addressed. According to the "Dynamic Monitoring Technical Guidelines for Soil and Water Loss" issued by China's Ministry of Water Resources in 2020, anthropogenically disturbed soil erosion parcels (ADPSE) are defined as land disturbed by human activities, such as construction and production, which may lead to soil erosion. Examples include quarries, sand and gravel extraction sites, brick kilns, spoil dumps, tailings storage sites, and ongoing construction projects. In recent years, with the rapid advancement of industrialization and urbanization in China, intensity in energy development and urban construction has been increasing. Consequently, the number of parcels affected by anthropogenic soil erosion due to construction and production activities has been on the rise, thereby exacerbating the overall trend of soil erosion [5,6].

Early monitoring of the dynamic changes in ADPSE primarily relied on field surveys supplemented by satellite observations, a method characterized by low timeliness and lengthy cycles [7]. With the advancement of remote sensing technology, the spatial and temporal resolution of remote sensing imagery has remarkably improved. High–resolution remote sensing images contain rich spatial information, such as texture, shape, and structure, making it possible to extract dynamic change information of targeted land features more accurately. Remote sensing change detection (CD) technology extracts change information by analyzing two or more images of the same area taken at different times. It is the only effective means for comprehensively monitoring the dynamic changes of land features [8]. Traditional machine learning approaches identify and extract information from multitemporal high–resolution remote sensing imagery, which contains rich surface details, by using manually designed features [9,10]. These methods include image differencing [11], principal component analysis (PCA) [12], change vector analysis (CVA) [13], random forest (RF) [14], and support vector machine–based techniques (SVM) [15]. Although the above methods achieve good recognition performance, their feature extraction relies on prior knowledge and parameter settings, which limits their broader applicability. Additionally, their ability to extract spatial features is relatively constrained, making it challenging to achieve optimal CD results on high–resolution images with complex spatial texture characteristics [9].

In recent years, deep learning has addressed the aforementioned issues in traditional machine learning and has become a popular trend in the field of remote sensing CD [16]. Depending on the stage of information fusion, these techniques can be broadly categorized into two types: single–stream and dual–stream architectures. The single–stream structure first fuses multitemporal data, which are then input into a deep neural network to highlight semantic information related to changes, guiding the network to learn change features from the imagery and converting the CD task into a classification problem. The dual–stream structure extracts deep features of bitemporal images through two network branches and subsequently fuses these features to extract change information, but this approach is mostly applied to detect change areas without distinguishing between change types [17,18]. Given that the change types in ADPSE are crucial for understanding soil erosion dynamics, this study adopts a single–stream network to transform the CD task into a classification task for experimentation.

In the field of CD, single–stream network architectures have gained prominence for their efficiency and effectiveness. Chen et al. [19] enhanced the Deeplabv3+ model, a prime example of a single–stream network that improves semantic segmentation. By utilizing atrous separable convolution within an encoder–decoder structure, the model achieves refined segmentation results, particularly along object boundaries. This approach sets new benchmarks on the PASCAL VOC and Cityscapes datasets by optimizing accuracy and speed in segmentation tasks without the need for postprocessing. Zhang et al. [20] applied a cascaded fully convolutional network with a single–stream architecture to de-

tect water bodies from synthetic aperture radar images. The model effectively managed complex background and edge information in water body segmentation, demonstrating the robustness of single–stream architectures for CD tasks in challenging environments. Peng et al. [21] proposed a novel end–to–end CD method based on an efficient semantic segmentation encoder–decoder architecture called UNet++. In this approach, coregistered image pairs are first concatenated and fed into the modified UNet++ network, where global and fine–grained information are leveraged to generate feature maps with high spatial resolution. A multiside output fusion strategy is then employed to combine change maps from different semantic levels, resulting in a final change map with high precision. Song et al. [22] proposed a single–stream network called 3M–CDNet, which integrates a backbone network with deformable convolutions to enhance the model's capability in geometric transformation modeling. Additionally, a two–stage feature fusion strategy is applied to improve feature representation. The model achieved state–of–the–art performance on the LEVIR–CD and season–varying datasets.

These above–mentioned methods struggle to balance local detail extraction and global context modeling, often resulting in feature information loss. Furthermore, these methods lack the ability to establish long–range dependencies in high–resolution images, making it difficult to capture subtle differences in target features. Combined with frequent noise artifacts like salt–and–pepper effects, these factors result in suboptimal segmentation quality and inaccurate boundary delineation. Meanwhile, high–resolution remote sensing images, with high spatial resolution, complex image characteristics, and rich geometric textures, often face challenges such as spectral confusion (same spectrum, different objects) and object confusion (same object, different spectra). Additionally, ADPSE and other land features exhibit significant variations in shape, size, texture, and color across different regions, further complicating accurate analysis and classification.

In order to solve the –abovementioned problems, this study proposes an improved SegFormer model for monitoring dynamic changes in ADPSE by category. The main contributions of this study are as follows:

(1) A cross–scale boundary constraint (CSBC) is introduced into the encoding stage, which is constructed to fuse high–level and low–level semantic features through different attention modules. This fusion provides cross–guidance for the boundary extraction of dynamic change categories in ADPSE.
(2) The pseudoresidual multilayer perceptron (PR–MLP) module introduces two depthwise convolution layers with different dilation rates and padding, capturing more comprehensive contextual information across different spatial hierarchies to enhance the precision of feature extraction.
(3) In the decoding stage, a multiscale feature fusion (MSFF) module is introduced to address the issues of class imbalance and noise interference in small target categories. This module fuses semantic features from different levels in the encoding stage, reducing the loss of semantic information due to the simplicity of the SegFormer decoder and enhancing feature fusion and propagation.
(4) To address the class imbalance between target and non–target regions, which can lead to the learning process to become stuck in local minima of the loss function, causing the classifier to be heavily biased towards the background class, we combined the cross–entropy loss with the Dice loss. This combination was used to construct boundary and segmentation loss functions, which effectively improved both the classification boundary accuracy and the overall segmentation performance.

2. Study Area and Data

Hubei Province, located in the heart of China, extends from 29°01′53″N to 33°6′47″N and 108°21′42″E to 116°07′50″E, covering an area of 185,900 km^2. The province serves as a transitional zone between China's second and third topographical steps. It is surrounded by mountains to the north, west, and east, with flatlands in the central region (Figure 1). With Hubei's economic development, activities such as urbanization, infrastructure construction,

and mineral resource exploitation have intensified, progressively increasing anthropogenic disturbances and leading to considerable soil erosion. These activities have caused extensive damage to surface vegetation, reducing the land's resistance to erosion. The construction of development zones, roads, and mining is among the primary drivers of soil loss in the province. According to 2021 soil erosion monitoring data, Hubei currently experiences soil erosion over an area of 31,090.85 km^2, with severe erosion spanning 2905.21 km^2, making it one of the provinces most heavily affected by erosion in China.

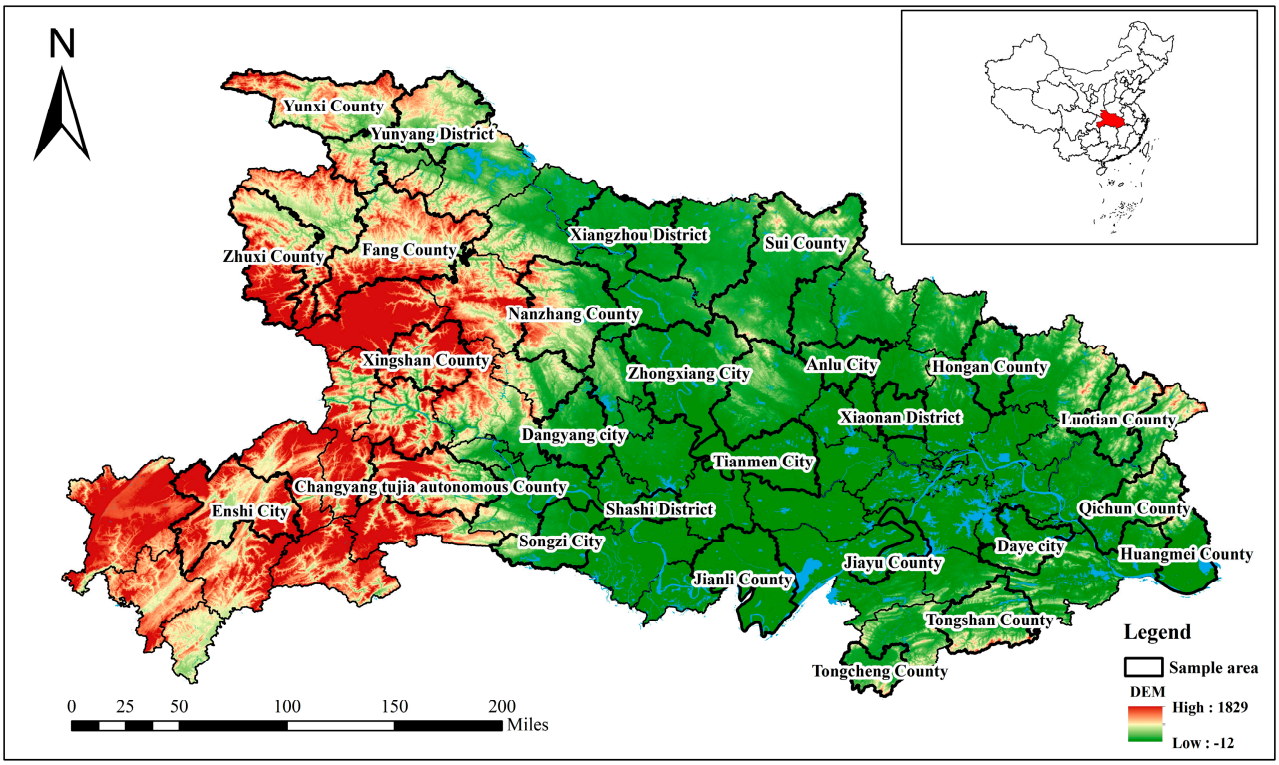


Figure 1. Distribution of sample areas in Hubei Province.

In this study, the dataset for changes in ADPSE was constructed using high–resolution imagery from the Gaofen–6 and Ziyuan–3 satellites, covering the period from January to May in 2020 and 2021. The imagery, with a fused resolution of 2 m, was stitched to create a comprehensive RGB mosaic of Hubei Province. The remote sensing imagery from the two time periods was combined into six–band data. Preprocessing steps were conducted, including image alignment, radiometric calibration, atmospheric correction, cropping, and filtering, to achieve more precise image information.

As shown in Figure 1, a combination of indoor classification monitoring and field verification was employed to obtain the distribution data of anthropogenically disturbed soil erosion parcels and other land cover types across 26 counties in Hubei over the two years. The nine land cover categories include anthropogenic soil erosion parcels, forest, dryland, water bodies, gardens, grassland, transportation land, paddy fields, and built–up land (data obtained from the Hubei Water Resources and Hydropower Research Institute). In the dataset, some transformations, such as ADPSE into water bodies, gardens, and paddy fields, as well as transitions from transportation land and gardens into ADPSE were minimal in quantity and thus excluded from the analysis.

As shown in Table 1, the study categorizes 12 dynamic change types for anthropogenic soil erosion parcels. The imagery and annotation data were clipped into 512 × 512pixel tiles. Following data cleaning, images in which background regions occupied more than 90% of the area were removed. A total of 9057 images were obtained, with 60% used for training, 20% for validation, and 20% for testing.

Table 1. Illustration of ADPSE change categories.

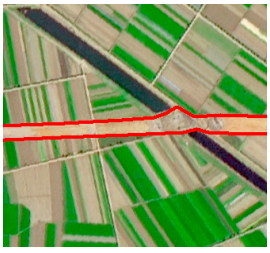
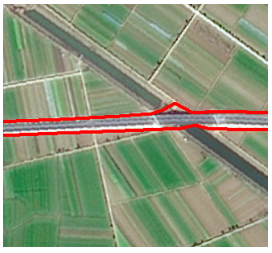
2020	2021	Sort	Image (2020)	Image (2021)
ADPSE	Forest	1		
ADPSE	Dryland	2		
ADPSE	Built–Up Land	3		
ASPSE	Grassland	4		
ADPSE	Transportation Land	5		
ADPSE	ADPSE	6		

Table 1. *Cont.*

2020	2021	Sort	Image (2020)	Image (2021)
Paddy field	ADPSE	7		
Grassland	ADPSE	8		
Water Bodies	ADPSE	9		
Built–Up Land	ADPSE	10		
Dryland	ADPSE	11		
Forest	ADPSE	12		

3. Materials and Methods

3.1. Network Structure of the Improved SegFormer

CD and classification in high–resolution remote sensing imagery have been prominent research focuses in the field of remote sensing. The complexity of scenes and diversity

of change targets in such imagery present considerable challenges to accurate CD and classification. To address these issues, this study proposes a CD semantic segmentation network based on a single–stream SegFormer architecture. By introducing improved feature extraction and fusion modules that combine semantic and change information, the model enhances the accuracy of change area detection and classification. This chapter provides a detailed description of the improved SegFormer network structure. As illustrated in Figure 2, the improved network builds upon the original backbone by incorporating the CSBC and MSFF modules and modifying the original multilayer perceptron (MLP) to enhance feature extraction and CD capabilities.

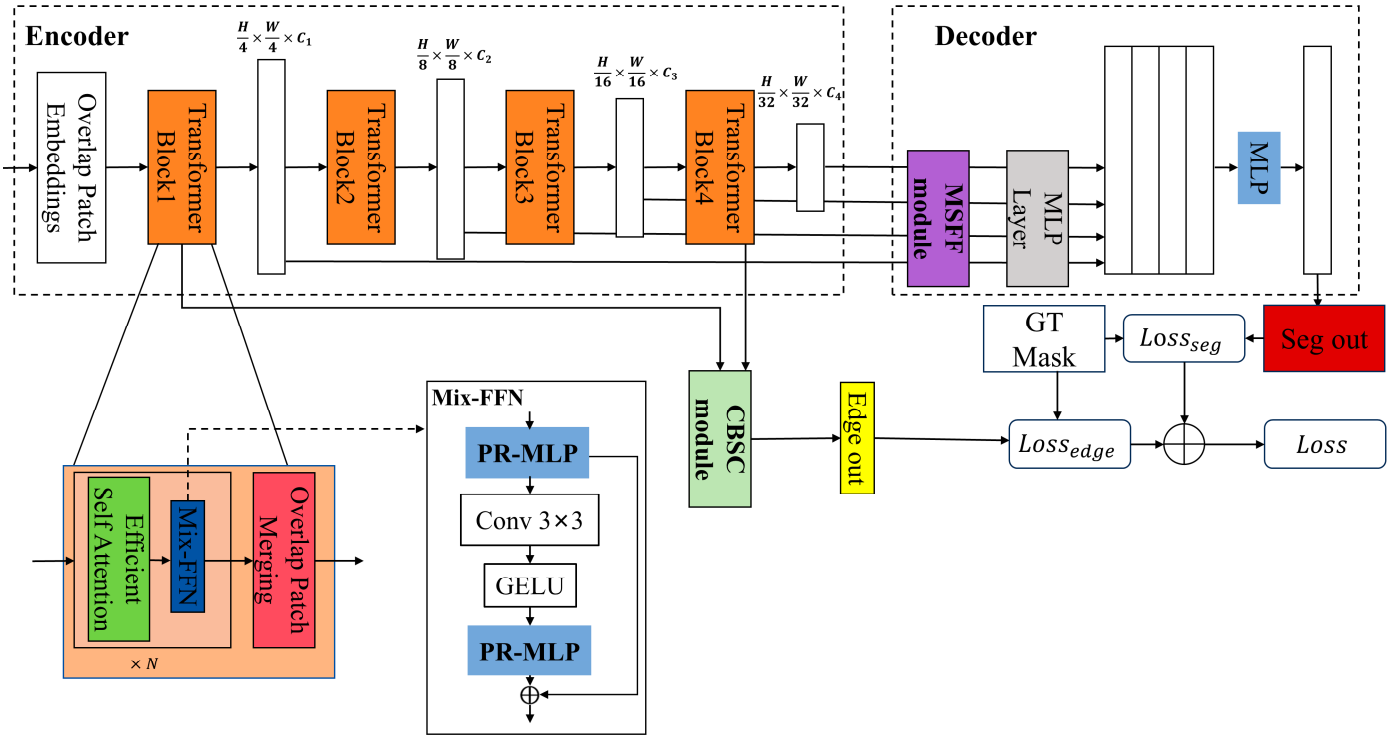


Figure 2. Network structure of the improved SegFormer.

3.2. Network Structure of the SegFormer

SegFormer effectively integrates a hierarchical Transformer encoder and an all–MLP decoder (Figure 3), achieving state–of–the–art performance in semantic segmentation tasks by efficiently handling multiscale features without positional encoding, thus enhancing robustness and adaptability to varying input resolutions while maintaining a compact model size. The SegFormer network is composed of three core components: overlap patch embedding (OPE), efficient multihead self–attention (EMSA), and mix feed–forward network (Mix–FFN). The OPE module slices the input data to extract and downsample features from the resultant image tensors, which are then processed through the EMSA and Mix–FFN layers to establish global contextual relationships. The network depth is incrementally enhanced by stacking multiple EMSA and Mix–FFN modules, facilitating the extraction of rich details and semantic features.

In the OPE module, the input image is decomposed into overlapping patches, which are then embedded to extract local spatial features effectively. This overlapping slicing technique retains more spatial information between adjacent patches, ensuring continuity at patch edges, thus enhancing the capability of the model to capture details of the input image. Specifically, for an input image with a resolution of $H \times W \times 3$, hierarchical feature maps Fi of dimensions $\frac{H}{2^{i-1}} \times \frac{w}{2^{i+1}} \times C_i$ are generated through successive patch merging operations, where $C_{i+1} > C_i$. Overlapped patch merging can transform a patch of $N \times N \times C_i$ into

a vector of $1 \times 1 \times C_{i+1}$, allowing the feature dimensions to reduce from $\frac{H}{4} \times \frac{w}{4} \times C_1$ to $\frac{H}{8} \times \frac{w}{8} \times C_2$ progressively.

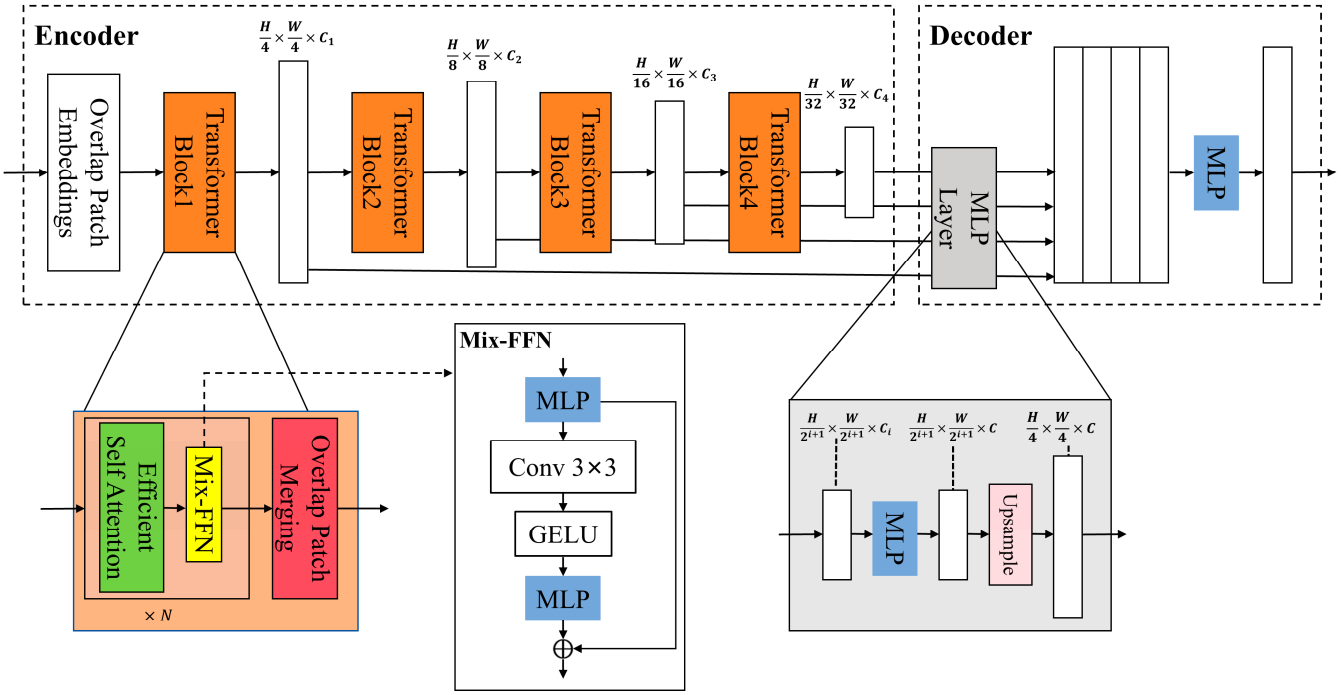


Figure 3. Network structure of SegFormer.

The primary role of the EMSA is to capture features across diverse spatial regions intricately by concurrently processing multiple attention heads. This methodology considerably augments the capability of the model to discern features at various scales and positions within images, thereby enhancing overall interpretability and effectiveness. The foundational principle of conventional self–attention mechanisms is delineated as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_{head}}}\right)V, \tag{1}$$

where Q, K, and V are matrices of size $N \times C$ ($N = H \times W$), representing query, key, and value, respectively.

The key innovation of the EMSA lies in the introduction of a sequence compression method to enhance the efficiency and performance of the traditional multihead self–attention mechanism. Initially, the length of matrix K is reduced from N to $\frac{N}{R}$ through a reshaping operation while simultaneously increasing the dimensionality of each element to preserve information. A linear layer then transforms the dimensions of the compressed K from $C{\cdot}R$ back to C, adjusting the dimensions accordingly. This step maintains the dimensionality of K while reducing the sequence length, as illustrated below:

$$\hat{K} = Reshape\left(\frac{N}{R}, C, R\right)(K) \tag{2}$$

$$K = Linear(C{\cdot}R, C)\left(\hat{K}\right), \tag{3}$$

where K represents the reduced sequence, and R is the reduction ratio.

Mix–FFN is an enhanced feedforward network utilized after each self–attention module, primarily aimed at further processing and refining features. In contrast to traditional Transformers that employ only linear layers in FFNs, Mix–FFN incorporates a 3×3 convolution and MLP operation within the FFN. This inclusion of a convolutional layer allows

the model to process local features more effectively and enhance the expression of positional information. The specific operation is detailed below:

$$x_{out} = MLP(GELU(Conv_{3\times3}(MLP(x_{in})))) + x_{in}, \quad (4)$$

where x_{out} represents the features passed from the self–attention module to the Mix–FFN (Feed–Forward Network). The activation function used in this process is GELU.

The Transformer, leveraging the self–attention mechanism, can process global information in a singular operation, thus exhibiting a notably larger effective receptive field compared with convolutional neural network (CNN) of equivalent depth. This architecture effectively captures extensive contextual information at various levels, providing the decoder with rich, multiscale features. Consequently, the SegFormer network simplifies decoding by employing solely MLP layers, considerably reducing computational demands and model parameters. During the decoding phase, input features are initially processed through an MLP layer. Subsequently, these features are dimensionally normalized through a specifically designed linear layer (Equation (5)). The processed features are then upsampled to match the resolution of the initial encoder block via bilinear interpolation techniques (Equation (6)). Upon completion of the upsampling process, features across different layers are concatenated along the channel dimension to integrate varied level features (Equation (7)); ultimately, the concatenated features undergo final semantic integration through an additional linear processing layer, culminating in the output of the final semantic segmentation results (Equation (8)).

$$\hat{F}_i = Linear(C_i, C)(F_i)\ \forall i, \quad (5)$$

$$\hat{F}_i = Upsample\left(\frac{W}{4} \times \frac{W}{4}\right)(\hat{F}_i)\ \forall i \quad (6)$$

$$F = Linear(4C, C)(Concat(F_i)),\ \forall i \quad (7)$$

$$M = Linear(C, N_{cls})(F) \quad (8)$$

The SegFormer network utilizes Mix Transformer (MiT) B0–B5 as its backbone for feature extraction, with model configurations increasing in parameters and computational complexity from B0 to B5. Given the importance of accurately identifying dynamic changes in ADPSE for policymaking in soil and water conservation, MiT–B5 was selected to maximize recognition accuracy. The hyperparameters for MiT–B5 are shown in Table 2.

Table 2. MiT–B5 hyperparameters.

Name	Number
Embed dims	64
Num layers	[3, 6, 40, 3]
Num heads	[1, 2, 5, 8]
Patch size	[7, 3, 3, 3]
Strides	[4, 2, 2, 2]
Sr ratios	[8, 4, 2, 1]
MLP ratio	[4, 4, 4, 4]

3.3. Structure of Pseudo–Residual Multilayer Perceptron

Section 3.2 mentions that the Mix–FFN used in SegFormer includes an FFN composed of 3 × 3 convolution and MLP operations, allowing the model to process local features more effectively and enhance positional information. However, the MLP structure of Mix–FFN is relatively simple, primarily relying on linear layers and a depthwise convolution (DWConv), which can only extract features within a fixed receptive field.

Given the variety of dynamic change categories in anthropogenic soil erosion parcels and the varying feature scales across these categories, this single–scale feature extraction

may lead to insufficient local information for some categories, requiring multiscale feature extraction. To address this, the PR–MLP introduces two depthwise convolution layers with different dilation rates and padding, enabling the capture of features at multiple scales. This improvement allows the model to handle information of varying spatial sizes flexibly, enhancing the perception of local and global information. During the encoding stage of SegFormer, each block contains a different number of MLPs. Given that MiT–B5 is used as the backbone network for feature extraction, the network's depth is important, potentially leading to gradient vanishing issues and suboptimal training performance. As network depth increases, information may not be efficiently transmitted, and simple layer stacking can hinder smooth information flow within the network. Thus, residual connections are introduced in PR–MLP to mitigate the issue, allowing direct information propagation between layers. This direct propagation alleviates the risk of gradient vanishing and promotes efficient information flow, improving the training efficacy and convergence speed of the model. The PR–MLP can be instantiated as shown in Figure 4.

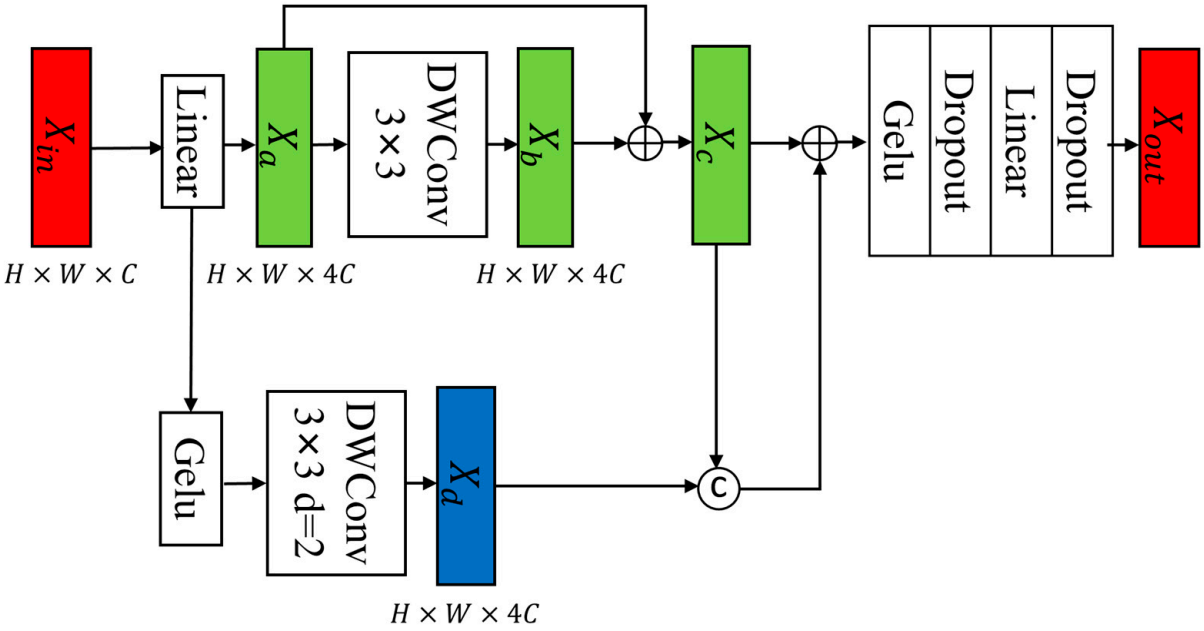


Figure 4. Structure of the PR–MLP.

The input feature X_{in} is processed through two depthwise convolution layers with different dilation rates and padding to capture features at various scales, producing features X_b and X_d. Feature X_a is then combined with X_b through a residual connection to form X_c. After concatenating X_c and X_d, the resulting feature is further fused with X_c to obtain the final feature representation. These steps address the limitations of the original MLP, including insufficient feature representation, gradient vanishing, poor information flow, and the risk of overfitting, remarkably improving the model's performance and generalization capability. The calculations for X_a, X_b, X_c, X_d, and X_{out} are detailed in Equations (9)–(13).

$$X_a = Linear(X) \tag{9}$$

$$X_b = DWConv_{3\times3}(X_a) \tag{10}$$

$$X_c = X_a + X_b \tag{11}$$

$$X_d = DWConv_{3\times3}^{d=2}(Gelu(Linear(X))) \tag{12}$$

$$X_{out} = Dropout(Linear(Dropout(Gelu(X_c + (concat(X_c, X_d)))))) \tag{13}$$

3.4. Structure of CSBC

In computer vision, high–level and low–level features represent different levels of information in image processing. Low–level features have higher resolution and contain more boundary and detail information. However, given fewer convolutional layers being applied, they exhibit lower semantic richness and higher noise. In contrast, high–level features possess stronger semantic information, but their lower resolution limits their ability to perceive fine details [23]. A complementary relationship exists between high–level and low–level features, and therefore, in this module, high– and low–level features are fused to guide boundary prediction. The structure of the CSBC module is illustrated in Figure 5.

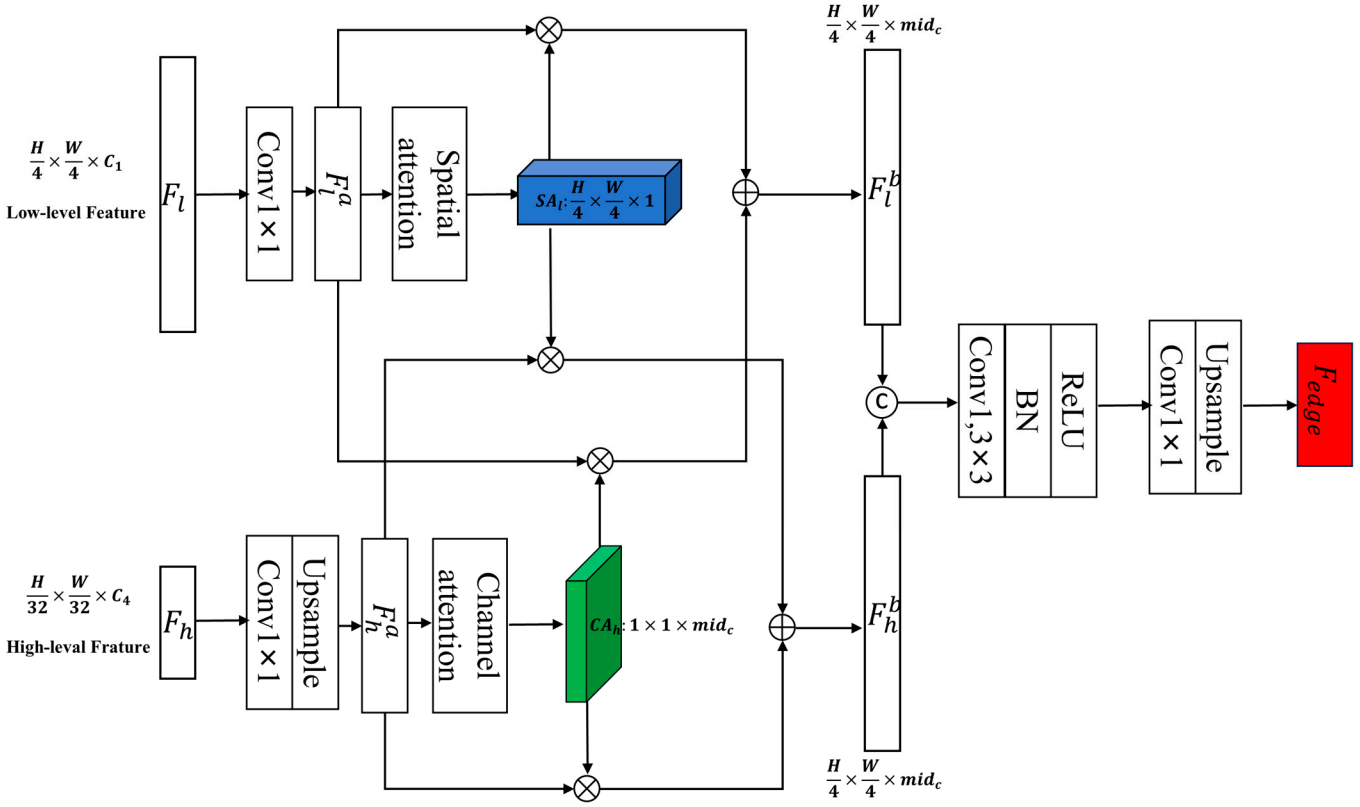


Figure 5. The structure of the CSBC.

First, low–level and high–level features, denoted as F_l and F_h, undergo channel mapping through conv 1 × 1, resulting in a new intermediate number of channels mid_c, where $C_1 < mid_c < C_4$. In this study, mid_c is set to 96. This operation increases the number of features in F_l, enhancing its semantic information while reducing the dimensionality of F_h, thereby decreasing computational complexity.

Given that low–level features contain rich spatial information and high–level features hold abundant channel information, spatial and channel attention mechanisms [24] are introduced to improve boundary prediction accuracy. Low–level features, denoted as $F_l^a \in \left(mid_c, \frac{H}{4}, \frac{W}{4}\right)$, are passed through a spatial attention module to enhance the representation of key regions, resulting in spatial attention weights SA_l. High–level features, denoted as $F_h^a \in \left(mid_c, \frac{H}{4}, \frac{W}{4}\right)$, undergo channel attention to strengthen important feature channels, producing channel attention weights CA_h.

To obtain more accurate boundary features and enable cross–guidance between high– and low–level features, F_l^b and F_h^b are calculated as shown in Equations (14) and (15).

Finally, F_l^b and F_h^b are fused and further processed through convolution to learn the features, yielding the boundary prediction result F_{edge}, as shown in Equation (16).

$$F_l^b = SA_l(conv_{1\times1}(F_l)) + CA_h(conv_{1\times1}(F_l)) \tag{14}$$

$$F_h^b = CA_h(Upsample(conv_{1\times1}(F_h))) + SA_l(Upsample(conv_{1\times1}(F_h))) \tag{15}$$

$$F_{edge} = Upsample\left(conv_{1\times1}\left(Relu\left(BN\left(conv_{3\times3}\left(concat\left(F_l^b, F_h^b\right)\right)\right)\right)\right)\right) \tag{16}$$

3.5. Structure of MSFF

The original SegFormer's decoding layer simply upsamples and concatenates the outputs from the four different Transformer modules without adequately integrating the multiscale feature maps. This basic concatenation and upsampling not only fail to effectively restore the features extracted by the encoding layer but can also lead to the loss of key target feature information. Additionally, the MLP decoder in the original SegFormer processes multiscale features by directly mapping feature maps to different channel dimensions, without detailed coupling across the channel and spatial dimensions. This results in the lack of fine–grained contextual modeling that combines spatial and channel information.

To further restore image information, this study proposes an MSFF module. By aligning features through upsampling and downsampling across different levels, the module progressively fuses multiscale features. Moreover, it incorporates channel and spatial attention mechanisms to mitigate the feature information loss caused by direct upsampling while effectively integrating information at different scales. This enables unified processing of global and local information, thereby enhancing feature propagation and fusion. The structure of the MSFF module is illustrated in Figure 6.

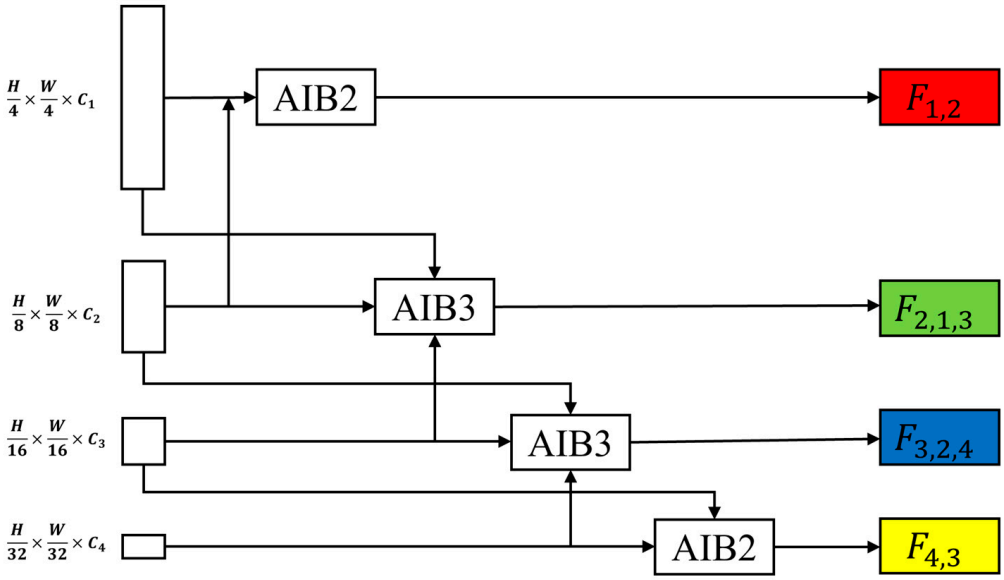


Figure 6. Structure of the MSFF.

The MSFF module first aligns the spatial dimensions of feature maps from different levels through upsampling and downsampling operations. After the features are aligned, the MSFF module concatenates these multiscale feature maps along the channel dimension. To process and fuse these concatenated feature maps, the MSFF module employs an attention–infused block (AIB). The AIB module consists of convolutional layers, channel attention, and spatial attention mechanisms, which further optimize and weight the fused features. The structure of the AIB is illustrated in Figure 7.

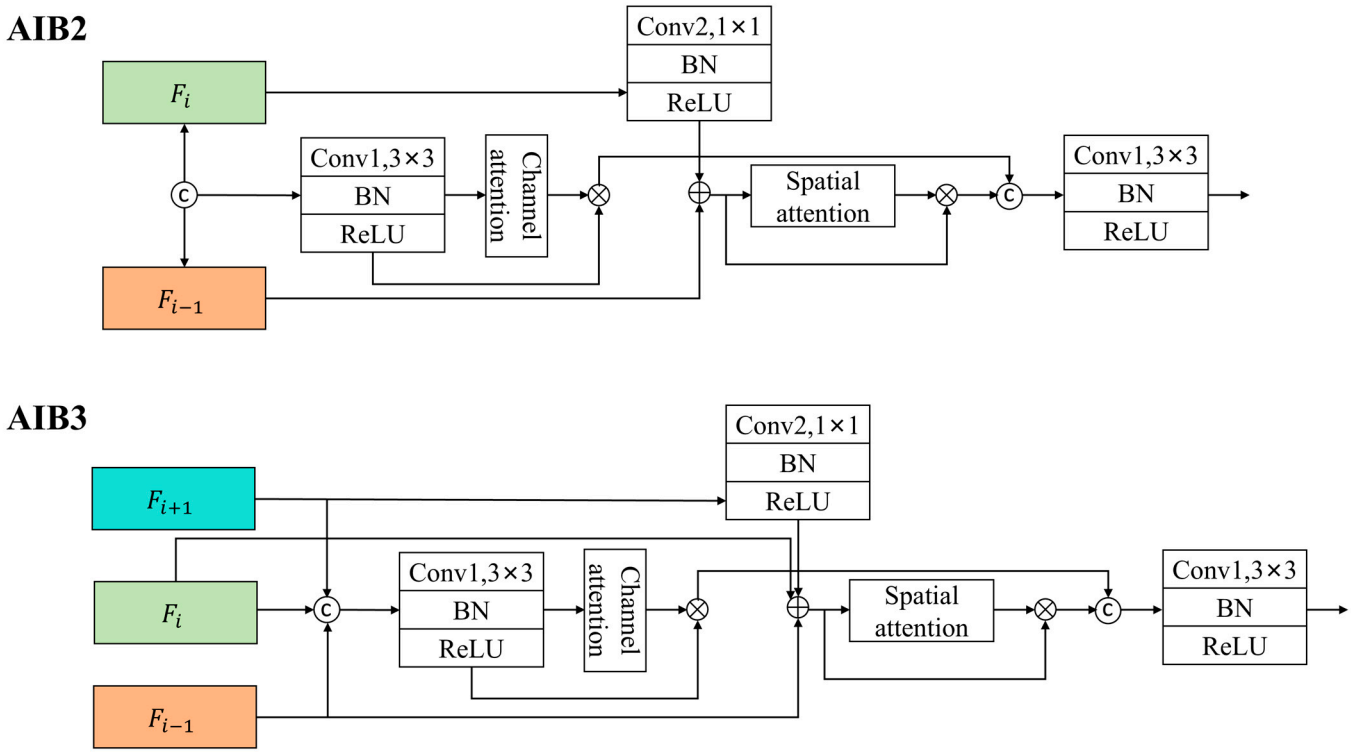


Figure 7. Structure of the AIB.

3.6. Loss Function

In deep learning–based image segmentation tasks, the loss function is a critical tool that guides the model's learning process. Traditional loss functions are often used to measure overall pixel classification accuracy. However, they exhibit considerable limitations when class imbalance exists in the dataset, particularly when the target boundary regions are complex. Given that the majority of pixels belong to the background class, the model tends to focus on optimizing the classification accuracy of the background region while neglecting the small, yet crucial target boundaries. This can result in inaccurate predictions in boundary regions, leading to blurry or mis–segmented edges in the final segmentation results.

An edge loss was constructed by combining cross–entropy loss [25] and Dice loss [26] to address this issue. In contrast to traditional loss functions, edge loss specifically targets the optimization of different target boundary regions. By increasing the emphasis on boundary pixels, the model is encouraged to improve the accuracy of target edges during training. The hybrid loss function is defined as follows:

$$Loss_{Total} = Loss_{Edge} + Loss_{Seg} \tag{17}$$

$$Loss_{Edge} = 1 - \left(\sum_{i=1}^{N} \sum_{c=1}^{C} e_{\text{target}}^{(i,c)} log(p_n^{(i,c)}) + \frac{2\sum_{i=1}^{N} p_n^i e_{\text{target}}^i}{\sum_{i=1}^{N} p_n^i + \sum_{i=1}^{N} e_{\text{target}}^i} \right) \tag{18}$$

$$Loss_{Seg} = \sum_{j=1}^{n-1} \left(- \sum_{i=1}^{N} \sum_{c=1}^{C} y^{(i,c)} log(p_j^{(i,c)}) + \left(1 - \frac{2\sum_{i=1}^{N} p_j^i y^i}{\sum_{i=1}^{N} p_j^i + \sum_{i=1}^{N} y^i} \right) \right), \tag{19}$$

In the loss function $Loss_{Edge}$, $e_{\text{target}}^{(i,c)}$ represents the label indicating the predicted boundary of the i pixel belonging to class c, and $p_n^{(i,c)}$ denotes the probability predicted by the model that the i pixel belongs to class c (output from the final layer N). Here, C and N represent the number of classes and the total number of pixels, respectively. In the segmen-

tation loss $Loss_{Seg}$, p_j^i refers to the probability of the target predicted at the i pixel in the j layer, and y^i is the true label of the i pixel.

3.7. Implementation Details

At the start of training, the model's weights are initialized randomly. Using a large learning rate at this stage can lead to instability in the model. This instability is addressed by employing a small warm–up learning rate during the initial epochs or steps. In this work, we applied the warm–up strategy for the first five epochs, followed by the "poly" learning rate strategy. The learning rate is then calculated using the following formula:

$$lr = base_lr \times \left(1 - \frac{epoch}{number_epoch}\right)^{power} \tag{20}$$

The model training parameters were configured as follows: the initial learning rate was 0.01, with a power value of 0.9. The total number of training epochs was set to 200, while the batch size during training was set to 12.

4. Experiments and Results

The experimental environment is CentOS 7.6. The running framework is PyTorch 2.1.2, and the processor is Intel(R) Core (TM) i9–12900 k CPU@3.20 GHz. The graphics card is NVIDIA A100, and the programming environment is Python 3.9.

4.1. Evaluation Indices

The network prediction results are evaluated using five metrics: intersection over union (IOU), recall, precision, F1–score (F1), and mean IOU (mIOU). IOU is defined as the ratio of the intersection to the union between the predicted values and the actual values for each class. Recall is the proportion of correctly predicted target features out of all actual target features, while precision is the proportion of correctly predicted target features out of all predicted target features. *mIOU* represents the average *IOU* across all classes. These five metrics are calculated using the following formulas:

$$IOU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{21}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{22}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{23}$$

$$F_1^i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{24}$$

$$mIOU = \frac{1}{C}\sum_i^C IOU_i, \tag{25}$$

where TP is the number of pixels correctly predicted as change class i; FP is the number of pixels incorrectly predicted as change class i; FN is the number of pixels incorrectly predicted as other classes and background, excluding class i; and C is the number of change sorts.

4.2. Comparison of Different Model Experimental Results

The improved model was compared with five mainstream semantic segmentation models on the test set to evaluate the performance of the ISegFormer model. These models include DANet [27], TransUNet [28], STANet [29], DTCDSCNet [30], and DSIFNet [17]. Ad-

ditionally, ablation experiments were conducted to verify the effectiveness of the modules within the improved model.

The performance of all six models was assessed on the test set, and Table 3 presents a comparison of the five evaluation metrics across 12 change categories for each model.

Table 3. Evaluation indices of the different methods.

Method	Indicators (%)	1	2	3	4	5	6	7	8	9	10	11	12	mIOU (%)
DANet	Recall	75.12	78.66	86.21	**89.23**	79.45	88.35	**88.31**	86.12	**73.33**	86.24	85.22	69.18	68.7
	Precision	73.28	77.39	88.75	85.43	78.21	89.61	79.01	85.43	69.02	83.56	85.71	65.94	
	F1	74.19	78.02	87.46	87.29	78.82	88.98	83.40	85.77	71.11	84.88	85.46	67.52	
	IOU	58.97	63.96	77.72	77.44	65.05	80.14	71.53	75.10	55.17	73.73	74.62	50.97	
TransUNet	Recall	76.29	**79.65**	87.02	86.37	82.44	89.32	86.93	86.11	72.36	**88.28**	**87.45**	68.36	69.21
	Precision	76.11	75.46	88.33	86.22	79.03	89.73	81.56	85.34	72.01	81.69	82.33	67.83	
	F1	76.20	77.50	87.67	86.29	80.70	89.52	84.16	85.72	72.18	84.86	84.81	68.09	
	IOU	61.55	63.26	78.05	75.89	67.64	81.03	72.65	75.01	56.47	73.70	73.63	51.62	
STANet	Recall	75.62	76.14	85.32	86.57	80.12	87.66	82.47	85.72	72.16	83.21	83.77	65.55	69.11
	Precision	**77.14**	77.29	89.53	84.22	81.61	90.21	**85.13**	**87.21**	**73.88**	84.55	85.12	**70.12**	
	F1	76.37	76.71	87.37	88.38	80.86	88.91	83.78	86.46	**73.01**	83.87	84.44	67.76	
	IOU	61.78	62.22	77.58	77.49	67.87	80.05	72.10	76.15	**57.49**	72.23	73.07	51.24	
DTCDSCNet	Recall	76.22	79.82	88.33	87.66	81.25	90.12	84.65	88.63	71.58	85.43	84.97	**71.24**	70.09
	Precision	73.47	74.43	89.02	87.37	81.99	91.33	83.03	86.88	73.22	87.21	85.06	66.47	
	F1	74.82	77.03	88.67	87.51	81.62	90.72	83.83	**87.75**	72.39	86.31	85.01	68.77	
	IOU	59.77	62.64	79.65	77.80	68.94	83.01	72.16	**78.17**	56.73	75.91	73.94	52.41	
DSIFNet	Recall	73.26	75.61	89.34	87.93	82.03	88.26	85.45	86.94	70.68	86.27	86.24	69.08	69.28
	Precision	75.69	75.12	88.26	86.88	83.00	92.11	83.21	84.36	67.55	83.45	84.55	69.71	
	F1	74.46	75.36	88.80	87.40	82.51	90.14	**84.32**	85.63	69.08	84.84	85.39	69.39	
	IOU	59.31	60.46	79.85	77.62	70.30	82.06	72.88	74.87	52.76	73.67	74.50	53.13	
ISegformer (Ours)	Recall	78.15	78.97	**91.35**	89.22	**82.56**	**94.41**	85.66	**88.64**	71.15	86.72	86.88	70.89	**72.34**
	Precision	**77.07**	**80.26**	**91.22**	**88.98**	**83.21**	**93.25**	84.12	86.37	73.45	**87.53**	**86.15**	69.21	
	F1	**77.61**	**79.61**	**91.28**	**89.10**	**82.88**	**93.83**	**84.88**	87.49	72.28	**87.12**	**86.51**	**70.04**	
	IOU	**63.41**	**66.13**	**83.97**	**80.34**	**70.77**	**88.37**	**73.74**	77.76	56.59	**77.18**	**76.23**	**53.89**	

As shown in Table 3, the bolded data represents the best performance achieved by this method in a certain category. ISegFormer achieved the best results in 10 of the 12 change categories, with IOU and F1–scores surpassing those of the other five models. In category 8, DTCDSCNet achieved an F1–score of 87.75% and an IOU of 78.17%, which are 0.26% and 0.41% higher, respectively, than those of our model. This result can be attributed to DTCDSCNet's effective use of a dual–attention network and feature pyramid, which allows it to better capture the features of change categories. Given limitations in receptive field size, our model's accuracy is slightly lower than DTCDSCNet in detecting large–scale grassland changes to ADPSE. In category 9, STANet achieved an F1–score of 73.01% and an IOU of 57.49%, outperforming our model by 0.73% and 0.90%, respectively. The reason is that STANet employs a spatiotemporal self–attention module that applies attention weighting to pixels within multiscale subregions, effectively capturing spatiotemporal dependencies across different scales. Consequently, it performs well even with smaller training datasets, while our model's accuracy decreases due to the limited quantity of this category in the dataset. ISegFormer, through the improved encoding and decoding mechanisms based on SegFormer, achieved an mIOU of 72.34% across the 12 change categories, outperforming the other models by a margin of 1.44–3.64%. This result indicates the model's deep understanding of features and its effective capability to integrate information. The high values of precision and recall also reflect the model's ability to capture edges and details accurately.

Among the 12 categories, category 6 had the highest accuracy, while category 12 had the lowest, with a difference of 1.44%. This result is mainly due to the imbalanced distribu-

tion of change categories in the dataset; ISegFormer performs better on categories with a higher representation in the data.

Four scenarios were randomly selected from the test set for prediction, and the prediction results of the six models are shown in Figure 8.

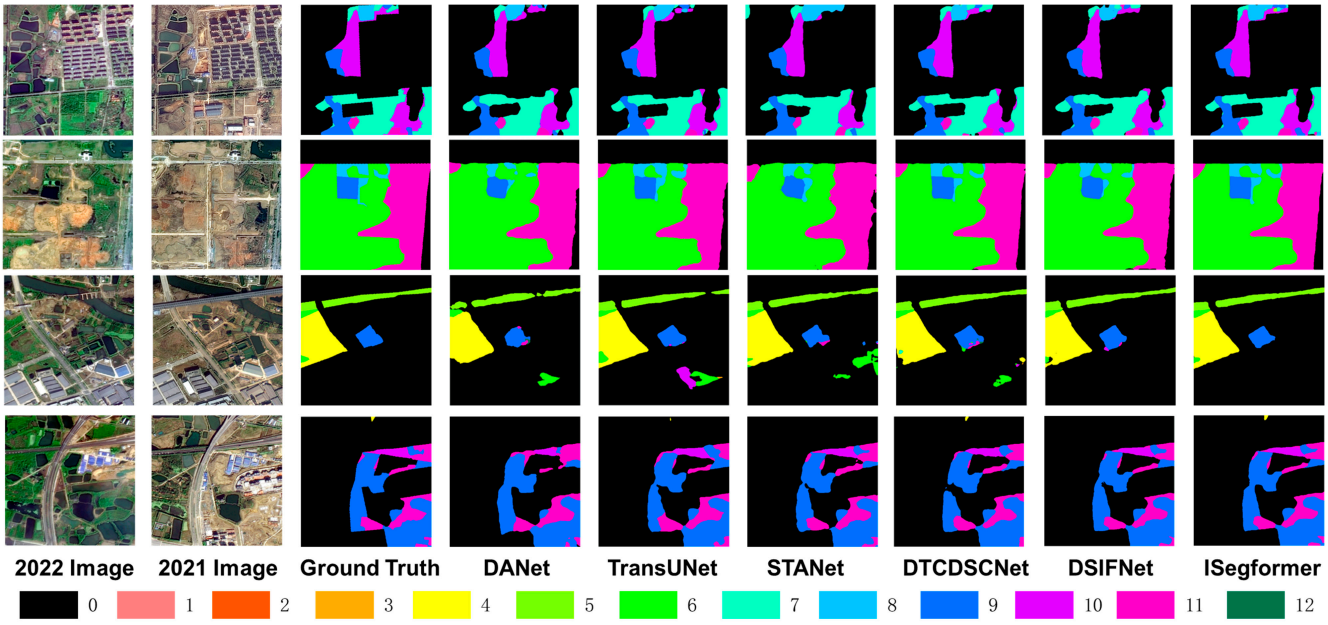


Figure 8. Segmentation results by different models on the test dataset.

In Scenario 1, the main changes are characterized by the transition from ADPSE to forest, dryland, construction land, grassland, and transportation land. The ISegFormer model demonstrates exceptional global recognition capability when segmenting these large–scale land–use type changes, accurately delineating the regions and shapes of land transitions. Specifically, for Categories 1 and 4, the presence of the CSBC module remarkably enhances the separability between categories and optimizes edge segmentation. STANet performs consistently in recognizing Categories 4 and 5, extracting spatiotemporal features with high accuracy in detecting land cover changes. By contrast, TransUNet shows a better advantage in handling Category 3 but encounters class fusion issues in Categories 2 and 4, where it struggles to distinguish vegetation from other land features clearly. DTCDSCNet and DSIFNet are proficient in segmenting large target regions but fail to capture small–scale details in the transformation of Category 5.

Scenario 2 involves changes from various land cover types to ADPSE. These transitions are diverse and regionally complex, requiring precise differentiation between classes. ISegFormer performs exceptionally well in Categories 9, 8, and 10, with its segmentation closely resembling ground truth, retaining the boundaries and overall shapes of these changing classes effectively. STANet exhibits strong CD capability in Categories 7 and grass–related Category 8 due to its spatiotemporal attention mechanism. DTCDSCNet shows advantages in Categories 11 and 12, especially in segmenting and recognizing large–scale regional changes with high stability. Although DANet and TransUNet are sensitive to large–area changes, they struggle with detailed segmentation in Categories 7–12, often affected by similar features across regions.

In Scenario 3, which showcases areas with no considerable change in ADPSE, ISegFormer and TransUNet maintain good stability in segmenting these unchanged regions, accurately preserving the integrity of unchanged land features. DTCDSCNet also displays robustness in recognizing large unchanged areas. However, STANet and DANet introduce some misclassifications along the edges of these unchanged regions, indicating their heightened sensitivity to subtle boundary changes, possibly due to noise interference.

Scenario 4 contains a mix of various change categories, exhibiting high complexity and diversity. In such diverse scenarios, ISegFormer demonstrates outstanding generalization ability and advantages in global feature fusion, effectively managing complex change relationships between land features while maintaining high classification accuracy and region segmentation integrity. Its multiscale feature fusion strategy enhances the model's performance in identifying land feature changes at different scales. STANet performs particularly well in processing Categories 1, 4, and 8, distinguishing different types of vegetation and their transformations. DTCDSCNet shows good specificity in Categories 3, 5, and 10, providing accurate and stable segmentation, especially in capturing land–use changes driven by human activities. However, TransUNet and DANet tend to misclassify boundaries and slight differences between categories in this complex scenario.

Each model displays unique strengths across different change categories and scenarios. ISegFormer stands out as a comprehensive and robust segmentation model due to its superior feature fusion and global context extraction capabilities, performing well in complex scenes and across various change categories. Other models also demonstrate advantages in specific change types and scenarios; for example, TransUNet's Transformer structure shows excellent segmentation ability in large–scale change regions, making it particularly suitable for detecting changes in buildings and extensive land areas. DTCDSCNet is well–suited for segmenting regions with noticeable human activity.

4.3. Ablation Experiment

Ablation experiments were conducted in this study to verify the effectiveness of each module. The ablation experiments involved creating a control variable group by removing one functional module to confirm its necessity. These experiments examined the impact of the PR–MLP, CSBC, and MSFF modules on the performance of the improved model and explored the role of each module within the network. With SegFormer as the baseline model, the three improvement measures were added step by step to form different combinations. Table 4 presents a comparison of evaluation metrics for each combination.

Table 4. Comparison on accuracy of different combinations.

Segformer (Baseline)	PR–MLP	CSBC	MSFF	mRecall (%)	mPrecision (%)	Macro–F1 (%)	mIOU (%)
√				81.49	80.19	80.83	67.83
√	√			82.51	80.76	81.63	68.96
√	√	√		81.76	83.26	82.52	70.24
√	√		√	83.34	81.86	82.59	70.35
√	√	√	√	**83.72**	**83.40**	**83.55**	**72.34**

After introducing the PR–MLP module, the model's mRecall improved by 1.02%, mPrecision increased by 0.57%, Macro–F1 rose by 0.8%, and mIOU increased by 1.13%. The PR–MLP enhances overall segmentation performance by incorporating dilated convolution branches to capture multiscale features, resulting in a modest improvement. Upon adding the CSBC module, mRecall improved slightly by 0.27%, while mPrecision saw a remarkable increase of 3.07%, Macro–F1 increased by 1.69%, and mIOU rose by 2.41%. The primary contribution of the CSBC module is enhancing boundary detection, which considerably boosts segmentation precision (mPrecision) and overall segmentation consistency (mIoU). When the MSFF module is added on top of PR–MLP, mRecall increases by 1.85%, mPrecision rises by 1.67%, Macro–F1 improves by 1.76%, and mIOU increases by 2.52%, reaching 70.35%. The MSFF module effectively enhances the model's sensitivity to different target scales by integrating multiscale features within the decoder.

Finally, with the incorporation of all three modules (PR–MLP, CSBC, and MSFF), the model achieved its highest performance across all metrics: mRecall improved by 2.23%, mPrecision by 3.21%, Macro–F1 by 2.72%, and mIOU by 4.51%. The combination of these modules demonstrates considerable complementarity in capturing multiscale

features, enhancing boundary information, and integrating features, which contributes to a comprehensive improvement in the overall segmentation performance of the model.

Figures 9 and 10 illustrate the segmentation results under different combination strategies. As shown in Figure 9, the baseline model performs well in segmenting large–scale, evident changes, particularly maintaining high integrity in the pink and orange regions. However, it struggles with detailed areas, especially with small–scale or complex boundary changes of land features, such as Categories 7 and 9, where remarkable errors are observed in comparison with the ground truth. The boundaries tend to be blurred, with frequent misclassification.

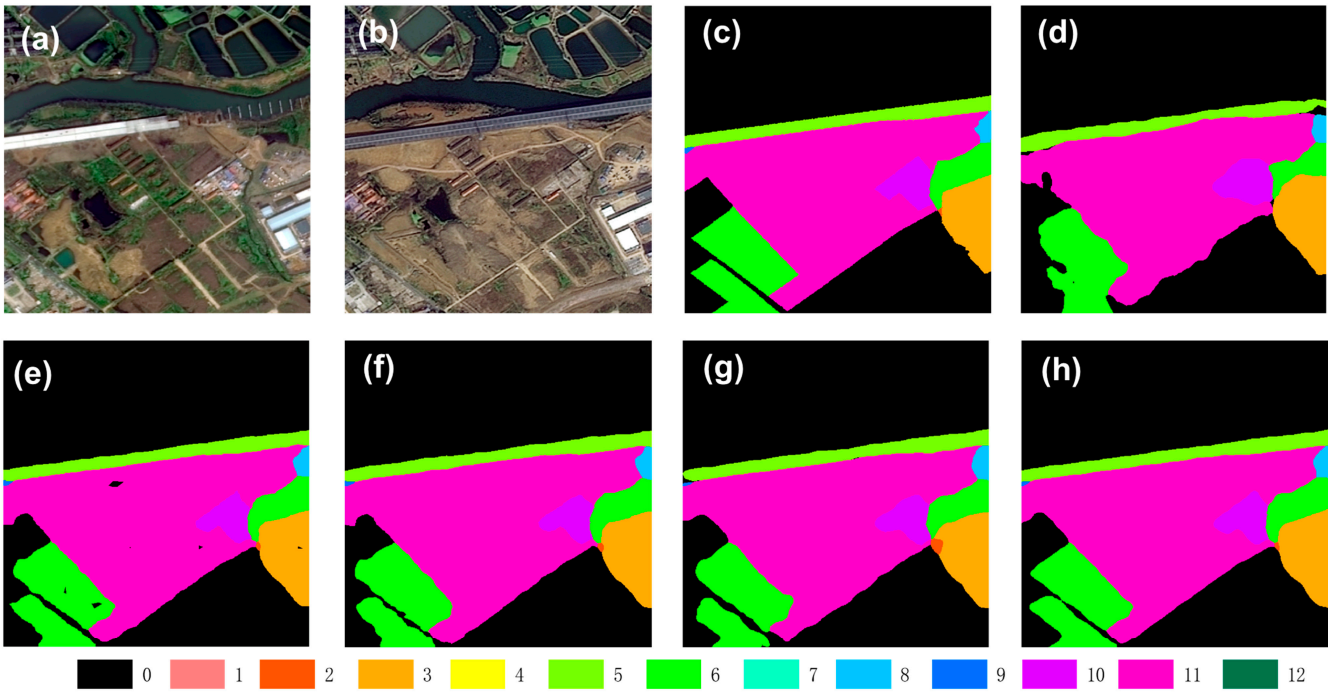


Figure 9. Ablation experiment segmentation results of roads under construction. (**a**) 2020 image. (**b**) 2021 image. (**c**) Ground truth. (**d**) Results predicted by SegFormer. (**e**) Results predicted by PR–MLP–SegFormer. (**f**) Results predicted by CSBC module added to PR–MLP–SegFormer. (**g**) Results predicted by MSFF module added to PR–MLP–SegFormer. (**h**) Results predicted by our improved model.

The PR–MLP module introduces dilated convolutions with varying dilation rates to capture multiscale features, resulting in remarkable improvements in boundary segmentation. For instance, the boundary between Category 6 (pink region) and Category 3 becomes more distinct, reducing confusion between unchanged and changed areas that the baseline model encounters. Additionally, it enhances detail capture for smaller regions, such as Category 5. The CSBC module further improves the segmentation of complex boundaries through boundary constraints. After adding the CSBC module, the model achieves more accurate boundary delineation, particularly for Categories 2 and 4, enhancing boundary smoothness and completeness. This module is especially effective in differentiating closely adjacent features of different categories, such as the subtle changes between Categories 5 and 6. When the MSFF module is added, the model's segmentation results remain stable across multiple regions. In particular, the consistency of segmentation boundaries improves remarkably in regions such as Categories 4 and 2, reflecting the MSFF module's stability and sensitivity to multiscale features.

By integrating all three modules—PR–MLP, CSBC, and MSFF—the model's ability to fuse multiscale features is remarkably enhanced. For large regions, such as Categories 2 and 4, the segmentation achieves optimal results with clear and complete boundaries. Simultaneously, the segmentation of small targets, such as in Categories 9 and 1, sees further optimization of boundary details. The combination of all three modules enables the

model to achieve an effective fusion of global and local features, resulting in segmentation outputs that most closely resemble the ground truth.

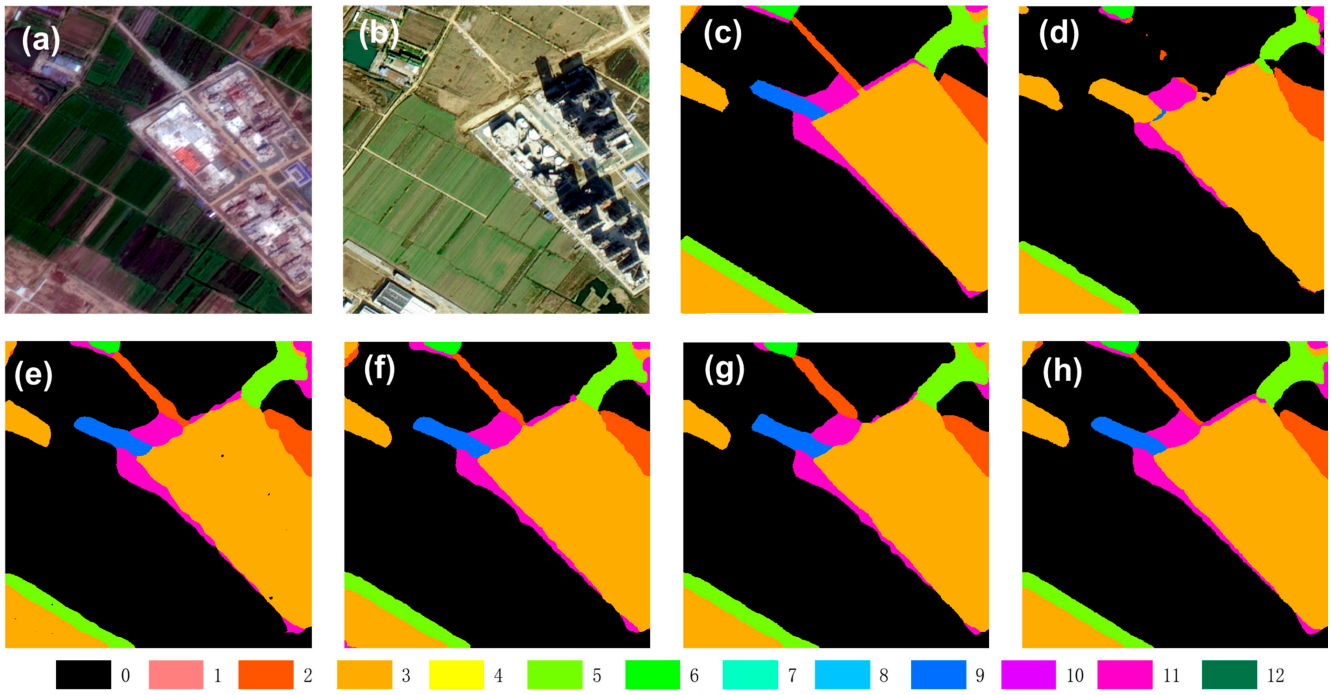


Figure 10. Ablation experiment segmentation results of roads under construction. (**a**) 2020 image. (**b**) 2021 image. (**c**) Ground truth. (**d**) Results predicted by SegFormer. (**e**) Results predicted by PR–MLP–SegFormer. (**f**) Results predicted by CSBC module added to PR–MLP–SegFormer. (**g**) Results predicted by MSFF module added to PR–MLP–SegFormer. (**h**) Results predicted by our improved model.

As shown in Figure 10, the original SegFormer model effectively segments large–scale, prominent land feature changes, maintaining a certain degree of overall coherence in the segmentation. However, it struggles with small–scale targets and complex boundary regions, such as in Category 9, where frequent misclassifications occur. After introducing the PR–MLP module, a notable improvement is observed in terms of identifying smaller targets, such as Category 1. Additionally, the boundary between Categories 5 and 2 becomes clearer with improved accuracy.

Upon adding the CSBC module, a comparison between Figures 10f and 10e shows remarkable enhancements in boundary processing for Categories 1 and 9, reducing the number of misclassifications. The CSBC module not only accurately segments large targets but also maintains precise recognition of complex boundaries.

When the MSFF module is integrated, the segmentation results in large land feature change regions, such as Categories 2 and 5, becoming more complete, with smoother boundary transitions. When all three modules (PR–MLP, CSBC, and MSFF) are incorporated, the model achieves optimal accuracy and completeness in segmentation.

Comparing Figure 10h with the earlier segmentation results reveals that the model demonstrates the ability to not only identify large–scale targets but also accurately segment small targets and regions with complex boundaries, ensuring detailed accuracy. The synergistic effect of the three modules enables the model to achieve optimal performance in multiscale feature capture, boundary constraint, and overall segmentation integrity.

In summary, the ablation experiments confirmed the effectiveness of each module and demonstrated the validity of the improvement strategies applied to SegFormer. The enhanced model considerably outperforms the original in segmentation accuracy and overall segmentation performance.

4.4. Dynamic Change Monitoring Analysis of ADPSE in Wuhan

The dynamic monitoring of ADPSE plays a crucial role in urban ecological management, sustainable land use, environmental risk prediction, informed decision–making, and smart city development. Therefore, this study utilized the improved ISegFormer model to map the dynamic change categories of ADPSE in Wuhan from 2020 to 2021 (Figure 11).

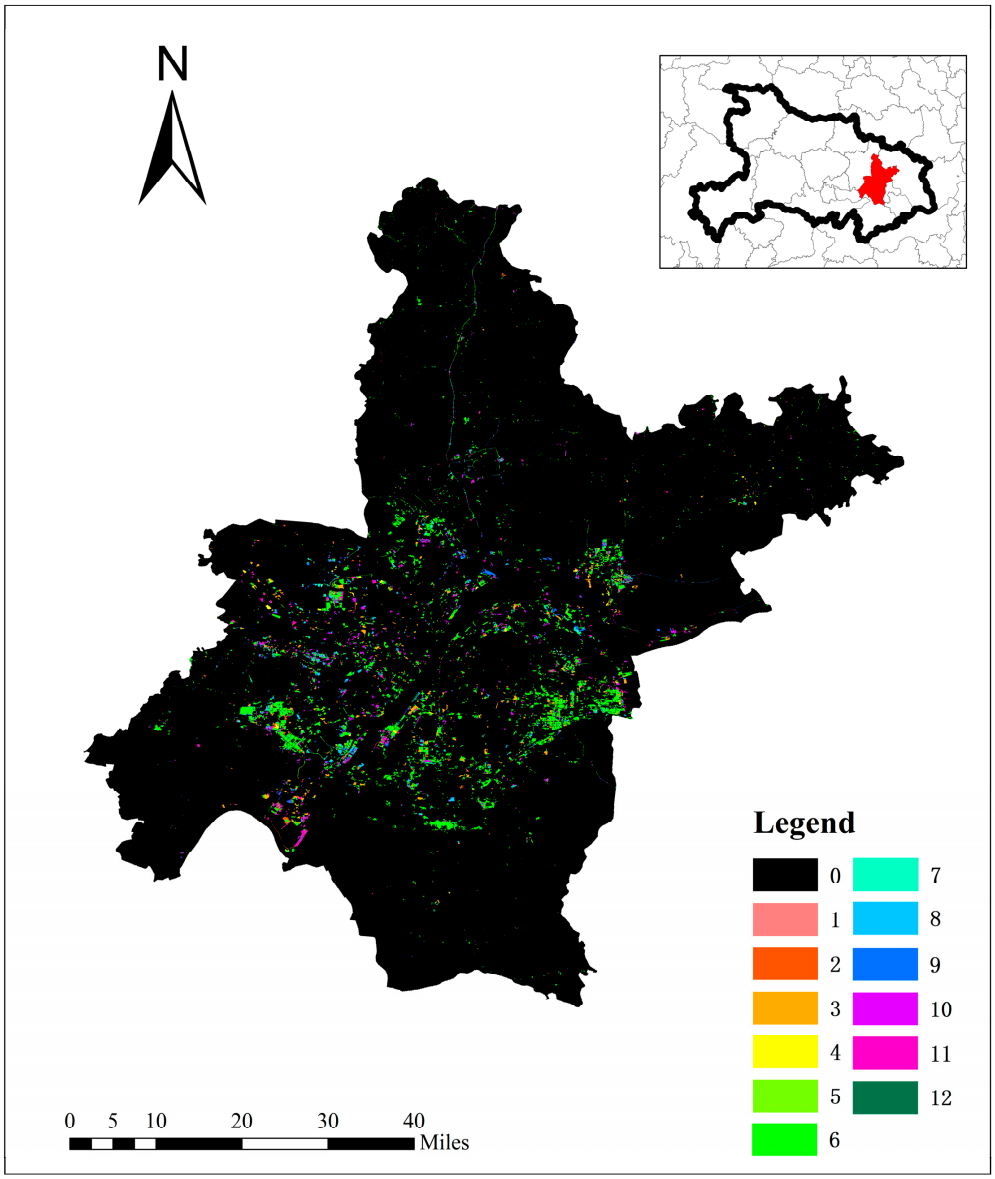


Figure 11. Spatial distribution of change categories.

In Figure 11, a large amount of orange indicates the transformation of ADPSE into construction land, with a total area of 32.21 km^2. This transformation is primarily located in Hongshan District and East Lake High–tech Zone, which are key areas of rapid urbanization and industrial development in Wuhan. These regions experience high land development intensity due to urban expansion and industrial park construction, resulting in the considerable conversion of ADPSE into construction land.

Huangpi District, located in the north of Wuhan, and Xinzhou District in the east serve as transitional zones between rural and urban areas, with notable agricultural and rural development activities. Categories 10 (construction land changing to ADPSE, 26.69 km^2) and 11 (dryland changing to ADPSE, 26.81 km^2) are widespread in Huangpi District. These changes are likely related to urban sprawl, construction projects, and land reclamation

in the area. Meanwhile, Categories 7 (paddy fields to ADPSE, 16.4 km^2), 8 (grassland to ADPSE, 18.25 km^2), and 9 (water bodies to ADPSE, 16.57 km^2) are predominantly found in Xinzhou District, associated with local agricultural development and hydrological engineering activities.

Caidian District and Jiangxia District, located in the southwestern and southern parts of Wuhan, contain many ecological protection areas, green spaces, and transportation networks. The figure shows that Categories 1 (ADPSE changing to forest, 2.21 km^2), 4 (ADPSE changing to grassland, 0.87 km^2), and 5 (ADPSE changing to transportation land, 4.77 km^2) are mainly concentrated in these areas. These changes reflect the active ecological restoration and greening efforts in Caidian and Jiangxia districts. At the same time, the expansion of transportation infrastructure has contributed to the conversion of some ADPSE into transportation land.

The most widespread category in the figure is Category 6 (no considerable change in ADPSE over two years), covering an area of 196.87 km^2. This phenomenon may be attributed to the relatively short time interval between the two remote sensing images, during which ADPSE did not undergo noticeable land–use changes. Generally, land–use changes require a long period for their effects to accumulate and become apparent. Therefore, over the short time span observed in this study, most areas retained their original state without considerable shifts in land characteristics.

5. Discussion

5.1. Model Chosen

In CNN models, the use of small–sized convolutional kernels in the initial stages tends to focus excessively on local image features, potentially overlooking critical global information. Although adding additional global context modules in the later stages of the model can partially mitigate this issue, the initial focus on small convolutional kernels can still adversely affect the segmentation performance. To address this problem, Dosovitskiy et al. [31] proposed the Vision Transformer (ViT) model, which incorporates a multilayer Transformer structure to handle image classification tasks by building global feature connections. Inspired by the ViT model, Zheng et al. [32] introduced SETR, which leverages the Transformer structure in each layer of the ViT encoder for global context modeling, and designed three different decoders specifically for image semantic segmentation tasks. Xie et al. [33] identified the limitation of ViT in outputting only single–scale features and proposed SegFormer, which outputs multiscale features while abandoning the position encoding used in ViT. This design choice avoids potential performance degradation when a difference in image resolution is found between training and testing phases. These models collectively demonstrate the high accuracy and robustness of single–stream networks in CD tasks. SegFormer simplifies the multi–scale encoder structure, reducing both computation and parameter count, while avoiding the need for complex positional encoding, thereby enhancing computational efficiency and training speed. Compared to traditional convolutional neural networks, it excels in capturing long–range dependencies and multi–scale features, demonstrating strong adaptability and achieving high–precision results across various segmentation tasks. Its efficiency, flexibility, and smaller parameter count make it the preferred choice for our base model.

5.2. Computation Costs

Increasing the model depth appropriately allows the model to extract deeper semantic features, thereby improving segmentation accuracy. However, as the depth increases, both the parameter count and computational complexity of the model also rise. Therefore, a trade–off between accuracy and complexity is necessary. Table 5 shows the comparison of the parameter count and computational complexity between the ISegformer model and other models.

According to Table 5, DANet exhibits the lowest parameter count of 72.61 MB, resulting in a relatively fast inference time of 1.78 ms, but its accuracy metrics, such as mIOU

(68.7%), fall behind other models. STANet and TransUNet have slightly higher parameter counts of 73.42 MB and 79.63 MB, respectively, with moderate inference times of 1.73 ms and 2.04 ms; however, their accuracy does not surpass ISegFormer in key metrics like mIOU or F1–score. DTCDSCNet, with the highest parameter count of 104.78 MB and an inference time of 2.59 ms, delivers robust performance in certain accuracy metrics but comes with a significant computational cost, making it less practical for resource–constrained applications. Similarly, DSIFNet strikes a balance between accuracy and computational cost, with 97.21 MB parameters and a 2.43 ms inference time, though its mIOU (69.29%) remains slightly lower than ISegFormer.

Table 5. Computation costs of different methods.

Method	Params (MB)	Inference Time (ms)
DANet	72.61	1.78
TransUNet	79.63	2.04
STANet	73.42	1.73
DTCDSCNet	104.78	2.59
DSIFNet	97.21	2.43
ISegformer(Ours)	88.07	1.89

ISegFormer, with 88.07 MB parameters and an inference time of 1.89 ms, demonstrates an efficient balance between computational load and performance. It outperforms other models in mIOU (72.34%) and key metrics like F1–score (91.28%) and Recall (94.41%) in specific categories. This highlights its ability to achieve superior accuracy without incurring excessive computational overhead. Compared with more computationally expensive models like DTCDSCNet and DSIFNet, ISegFormer significantly reduces complexity while maintaining competitive or superior performance across all evaluation metrics. Thus, ISegFormer achieves an ideal balance between accuracy and computational efficiency, making it a practical and high–performing choice for segmentation tasks.

6. Conclusions

Compared with other semantic segmentation models, the proposed ISegFormer model exhibits outstanding performance in monitoring the dynamic changes of ADPSE, particularly in the boundary recognition of complex land features and the segmentation of small target areas. By introducing PR–MLP, CSBC, and MSFF modules, the model has remarkably enhanced its capabilities in cross–scale feature extraction, boundary constraint, and multiscale feature fusion, achieving high segmentation accuracy and demonstrating strong generalization ability. The PR–MLP module effectively improves the model's perception of features at different scales, the CSBC module enhances the accuracy of boundary processing, and the MSFF module strengthens the multiscale fusion capacity of the decoder. The combination of these three modules allows ISegFormer to achieve comprehensive and accurate recognition of various land feature changes while being capable of handling feature changes across different spatial regions.

The conclusions of this study are as follows:

(1) The PR–MLP module introduces dilated convolution branches into the original MLP structure, using different dilation rates and padding strategies to capture multiscale features. Ablation experiments show that the introduction of the PR–MLP module remarkably improves the recall rate and mIOU of the model in segmenting complex and multiscale land features, with increases ranging from 1.13% to 2.52%. This module optimizes segmentation performance for large–scale and small–target features, making full use of multiscale information in remote sensing imagery.

(2) The CSBC module focuses on enhancing the ability to detect and recognize land feature boundaries by constructing a cross–scale boundary constraint mechanism. This module improves the model's processing of complex boundary information. When added to the baseline model with the PR–MLP module, CSBC boosts mPrecision and

Macro–F1 values by 1.67–3.07% and 1.69–2.72%, respectively, remarkably enhancing the fine–tuning of boundary segmentation. The CSBC module is particularly effective in detecting blurred boundary regions in practical segmentation tasks, ensuring edge continuity and accuracy in the segmentation results.

(3) The MSFF module effectively strengthens the model's ability to perceive and recognize changes at different scales by introducing a multiscale feature fusion mechanism into the decoder. This module remarkably improves segmentation coherence across various targets, particularly enhancing the overall coherence of segmentation results in large–area target regions. When added to the PR–MLP module, MSFF increases the model's mRecall by 1.85%, further enhancing boundary processing and target recognition capabilities, leading to a more comprehensive detection of land feature changes.

(4) ISegFormer, by integrating PR–MLP, CSBC, and MSFF modules, achieved optimal performance in the land feature change segmentation task using remote sensing imagery of Wuhan. Compared with other models (DANet, TransUNet, STANet, DTCDSCNet, and DSIFNet), ISegFormer shows considerable advantages across multiple evaluation metrics, with average increases of 2.72% in mIOU and 3.64% in Macro–F1. The model demonstrates stability and high–precision segmentation performance in detecting large–scale targets and regions with complex small–scale boundaries.

Despite the excellent performance of ISegFormer in detecting land feature changes, there is still room for improvement in detecting rapidly changing targets over short periods and enhancing the model's ability to recognize imbalanced categories. Future work may involve introducing higher temporal resolution remote sensing data, extending the observational period, and integrating multisource prior information or multispectral data to optimize model performance in complex scenarios.

Author Contributions: Conceptualization, Z.L. (Zhenqiang Li) and J.L. (Jialin Li); methodology, J.L. (Jialin Li); software, Z.L. (Zhenqiang Li); validation, J.L. (Jie Li); formal analysis, Z.L. (Zhangxuan Li); investigation, K.J.; resources, C.H.; data curation, Y.M.; writing—original draft preparation, J.L. (Jialin Li); writing—review and editing, J.L. (Jialin Li); visualization, Z.L. (Zhenqiang Li); supervision, Z.L. (Zhenqiang Li); project administration, J.L. (Jie Li); funding acquisition, Z.L. (Zhenqiang Li), J.L. (Jie Li) and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by grants from the National Nature Science Foundation of China (NSFC) Program (No. 42471517; No. 42401508; No. 42301527).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to their integration within the team's internal database, which is structured for internal access and use. Direct public access to this database is not feasible due to operational, privacy, and security protocols.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Telles, T.S.; Guimarães, M.D.F.; Dechen, S.C.F. The costs of soil erosion. *Rev. Bras. Cienc. Solo* **2011**, *35*, 287–298. [CrossRef]
2. Nearing, M.A.; Xie, Y.; Liu, B.; Ye, Y. Natural and anthropogenic rates of soil erosion. *Int. Soil Water Conserv. Res.* **2017**, *5*, 77–84. [CrossRef]
3. Thaler, E.A.; Kwang, J.S.; Quirk, B.J.; Quarrier, C.L.; Larsen, I.J. Rates of historical anthropogenic soil erosion in the midwestern United States. *Earth's Future* **2022**, *10*, 1–16. [CrossRef]
4. Huang, X.Z.; Ren, X.X.; Chen, X.M.; Zhang, J.; Zhang, X.S.; Shen, Z.W.; Hu, Y.; Chen, F.H. Anthropogenic mountain forest degradation and soil erosion recorded in the sediments of Mayinghai Lake in northern China. *CATENA* **2022**, *207*, 105597. [CrossRef]
5. Guo, L.J.; Liu, R.M.; Men, C.; Wang, Q.R.; Miao, Y.X.; Shoaib, M.; Wang, Y.F.; Jiao, L.J.; Zhang, Y. Multiscale spatiotemporal characteristics of landscape patterns, hotspots, and influencing factors for soil erosion. *Sci. Total Environ.* **2021**, *779*, 146474. [CrossRef]
6. Wang, L.; Yan, H.; Wang, X.W.; Wang, Z.; Yu, S.X.; Wang, T.W.; Shi, Z.H. The potential for soil erosion control associated with socio–economic development in the hilly red soil region, southern China. *CATENA* **2020**, *194*, 104678. [CrossRef]
7. Zhang, Y.; Zhou, Y.; Yao, H.; Sun, Y.; Yang, L.N.; Yang, X. Remote Sensing Monitoring and Development Discussion of Soil and Water Conservation for Construction Projects. *Chin. Flood Drought Manag.* **2021**, *31*, 163–167.

8. Zhu, Q.Q.; Guo, X.; Deng, W.H.; Shi, S.N.; Guan, Q.F.; Zhong, Y.F.; Zhang, L.P.; Li, D. Land–use/land–cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [CrossRef]
9. Shafique, A.; Cao, G.; Khan, Z.; Asad, M.; Aslam, M. Deep learning–based change detection in remote sensing images: A review. *Remote Sens.* **2022**, *14*, 871. [CrossRef]
10. Hecheltjen, A.; Thonfeld, F.; Menz, G. Recent advances in remote sensing change detection—A review. In *Land Use and Land Cover Mapping in Europ*; Braun, M., Ed.; Springer: Dordrecht, The Netherlands, 2014; Volume 18, pp. 145–178.
11. Ke, L.; Lin, Y.K.; Zeng, Z.; Zhang, L.F.; Meng, L.K. Adaptive change detection with significance test. *IEEE Access* **2018**, *6*, 27442–27450. [CrossRef]
12. Celik, T. Unsupervised change detection in satellite images using principal component analysis and *k*–means clustering. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 772–776. [CrossRef]
13. Liu, S.; Bruzzone, L.; Bovolo, F.; Zanetti, M.; Du, P. Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4363–4378. [CrossRef]
14. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support vector machine versus random forest for remote sensing image classification: A meta–analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
15. Wang, D.W.; Liu, S.W.; Zhang, C.; Xu, M.M.; Yang, J.F.; Yasir, M.; Wan, J.H. An improved semantic segmentation model based on SVM for marine oil spill detection using SAR image. *Mar. Pollut. Bull.* **2023**, *192*, 114981. [CrossRef]
16. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet–a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
17. Zhang, C.X.; Yue, P.; Tapete, D.; Jiang, L.C.; Shangguan, B.; Huang, L.; Liu, G.C. A deeply supervised image fusion network for change detection in high resolution bi–temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
18. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
19. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
20. Zhang, J.S.; Xing, M.D.; Sun, G.C.; Chen, J.L.; Li, M.Y.; Hu, Y.H.; Bao, Z. Water body detection in high–resolution SAR images with cascaded fully–convolutional network and variable focal loss. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 316–332. [CrossRef]
21. Peng, D.; Zhang, Y.; Guan, H. End–to–end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
22. Song, K.Q.; Cui, F.Z.; Jiang, J. An efficient lightweight neural network for remote sensing image change detection. *Remote Sens.* **2021**, *13*, 5152. [CrossRef]
23. Chen, Y.T.; Liu, L.W.; Tao, J.J.; Chen, X.; Xia, R.L.; Zhang, Q.; Xie, J.; Yang, K.; Xie, J.B. The image annotation algorithm using convolutional features from intermediate layer of deep learning. *Multimed. Tools Appl.* **2021**, *80*, 4237–4261. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
25. Mao, A.; Mohri, M.; Zhong, Y. Cross–entropy loss functions: Theoretical analysis and applications. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
26. Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; Li, J. Dice loss for data–imbalanced NLP tasks. *arXiv* **2019**, arXiv:1911.02855.
27. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
28. Chen, J.N.; Mei, J.R.; Li, X.H.; Lu, Y.Y.; Yu, Q.H.; Wei, Q.Y.; Luo, X.D.; Xie, Y.T.; Ehsan, A.; Wang, Y.; et al. TransUNet: Rethinking the U–Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
29. Chen, H.; Shi, Z. A spatial–temporal attention–based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
30. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual–task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [CrossRef]
31. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Zheng, S.X.; Lu, J.C.; Zhao, H.S.; Zhu, X.T.; Luo, Z.K.; Wang, Y.B.; Fu, Y.W.; Feng, J.F.; Xiang, T.; Philip, H.S.T.; et al. Rethinking semantic segmentation from a sequence–to–sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
33. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process Syst.* **2021**, *34*, 12077–12090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.